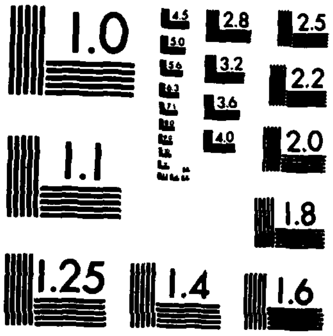
AD-A143 209 ROBUST MODEL SELECTION IN REGRESSION(U) PRINCETON UNIV 1/1
NJ DEPT OF STATISTICS E RONCHETTI FEB 84 TR-259-SER-2
ARO-19442.19-MA DAAG29-82-K-0178

UNCLASSIFIED F/G 12/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

2

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD-A143 209

# REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | | |
|---|---|---|
| 1. REPORT NUMBER<br>ARO 19442.19-MA | 2. GOVT ACCESSION NO.<br>N/A | 3. RECIPIENT'S CATALOG NUMBER<br>N/A |
| 4. TITLE (and Subtitle)<br>Technical Report No. 259 "Robust Model Selection in Regression" | | 5. TYPE OF REPORT & PERIOD COVERED |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Elvezio Ronchetti | | 8. CONTRACT OR GRANT NUMBER(s)<br>DAAG29-82-K-0178 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of Statistics<br>Princeton University<br>Princeton, N. J. 08544 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>U. S. Army Research Office<br>Post Office Box 12211<br>Research Triangle Park, NC 27709 | | 12. REPORT DATE<br>February 1984 |
| | | 13. NUMBER OF PAGES<br>10 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Akaike Information Criterion; $C_p$ criterion; M-estimators; Robust tests; Regression models.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A robust version of Akaike's model selection procedure for regression models is introduced and its relationship with robust testing procedures is discussed.

84 07 12 096

# Robust Model Selection in Regression

by

Elvezio Ronchetti

Technical Report No. 259, Series 2
Department of Statistics
Princeton University
February 1984

This work was supported in part by U.S. Army Research Office Grant Number DAAG29-82-K-0178.

# Robust Model Selection in Regression

by

Elvezio Ronchetti
Department of Statistics
Princeton University

## SUMMARY

A robust version of Akaike's model selection procedure for regression models is introduced and its relationship with robust testing procedures is discussed

*Some key words:* Akaike Information Criterion; $C_p$ criterion; M-estimators; Robust tests; Regression models.

## 1. INTRODUCTION

The Akaike Information Criterion is a powerful tool for choosing among different models that can be used to fit a given data set. If we denote by $L_p$ the log-likelihood of the model with p parameters, this amounts to choose the model that minimizes $-2L_p+2P$. This procedure may be viewed as an extension of the likelihood principle and is based on a general information theoretic criterion. In fact $2L_p-2P$ is a suitable estimate of the expected entropy of the model and by the Akaike Criterion the entropy will be, at least approximately, maximized; cf. Akaike (1973).

Bhansali and Downham (1977) proposed to generalize the Akaike Criterion by choosing the model that minimizes for a given fixed $\alpha$

$$AIC(p;\alpha) = -2L_p+\alpha\cdot p \ . \qquad (1)$$

Several proposals have been made for choosing $\alpha$; see, for instance, Bhansali and Downham (1977), Atkinson (1980). If we apply (1) to a linear regression model

$$y_i = x_i^T\theta + e_i \ , \qquad i=1,\dots,n \qquad (2)$$

with $n$ independent identically normally distributed errors with variance $\sigma^2$,

$$AIC(p;\alpha) = K(n,\hat{\sigma}) + R_p/\hat{\sigma}^2 + \alpha\cdot p \qquad (3)$$

where $K(n,\hat{\sigma})$ is a constant depending on the marginal of the $x_i$'s , $\hat{\sigma}^2$ is some estimate of $\sigma^2$ and $R_p = \sum_{i=1}^{n} (y_i - x_i^T \hat{\theta}_p)^2$ is the residual some of squares with respect to the least squares estimate $\hat{\theta}_p$ . AIC(p;2) is equivalent to Mallows' $C_p$ statistic; see Mallows (1973).

One of the main goals of robust statistics is to find new statistical procedures that are not influenced too much by small deviations from the distributional assumptions of the model. In recent years there has been a considerable amount of work directed to construct robust estimators and testing procedures for regression models, but the aspects related to a robust model choice have been somewhat neglected. Since the AIC statistic for regression models is a direct consequence of the normality assumption on the errors' distribution (see (3)), we cannot use it in this form with robust estimators and robust tests. The purpose of this note is to introduce a robust selection procedure for regression that, first, allows us to choose the model which fits the *majority* of the data taking into account that the errors might not be exactly normally distributed, and secondly, that can be used consistently with new robust estimators and tests.

In Section 2 the new robust procedure is introduced and its relationship to robust testing procedures is discussed. Section 3 presents some possible choices of the parameter $\alpha$ for the robust selection procedure.

## 2. A ROBUST SELECTION PROCEDURE

Let us assume that the errors in (2) follow some distribution with density g . Then the right hand side of (1) becomes

$$K(n,\hat{\sigma}) - 2 \sum_{i=1}^{n} \log g((y_i - x_i^T T_{n;p})/\hat{\sigma}) + \alpha \cdot p \ , \qquad (4)$$

where $T_{n;p}$ denotes the maximum likelihood estimator of $\theta$ when the errors' distribution is g . If we replace -log g in (4) by a general function $\rho$ , we obtain the following robust selection procedure. Note that a similar idea was used by Martin (1980) for autoregressive models.

For a given constant $\alpha$ and a given function $\rho$ , chooses the model that minimizes

$$AICR(p;\alpha,\rho) = 2 \sum_{i=1}^{n} \rho(r_{i;p}) + \alpha p \ , \qquad (5)$$

where $r_{i;p} = (y_i - x_i^T T_{n;p})/\hat{\sigma}$ , $\hat{\sigma}$ is some robust estimate of $\sigma$ and $T_{n;p}$ is the M-estimator defined as implicit solution of the system of equations

$$\sum_{i=1}^{n} \psi(r_{i;p}) x_i = 0 \ , \qquad (6)$$

with $\psi(r) = d\rho/dr$ .

The extension of AIC to AICR is the exact counterpart of that of maximum likelihood estimation to M-estimation; cf. Huber (1981, Section 3.2). In particular, if we choose $\rho$ as Huber's function

$$\begin{aligned} \rho_c(r) &= r^2/2 && \text{if } |r| \leqslant 0 \\ &= c|r| - c^2/2 && \text{otherwise ,} \end{aligned} \tag{7}$$

then $T_{n;p}$ is Huber's estimator and AICR $(p;\alpha,\rho_c)$ is the generalized Akaike statistic (1) computed under the least favorable errors' distribution with density

$$g_0(r) = (1-\varepsilon)(2\pi)^{-\frac{1}{2}}\exp(-\rho_c(r)) \ , \tag{8}$$

where $c$ is a function of the contamination $\varepsilon$ ; cf. Huber (1981, Chapter 4). In this case a robust estimate for $\sigma$ can be obtained using Huber's Proposal 2 (Huber 1981, p. 137) or Hampel's median absolute deviation (Hampel 1974, p. 388) in the model with all parameters.

Let us now investigate the relationship between AICR and robust testing procedures. Denote by $\theta^{(j)}$ the jth component of the vector $\theta$ and let

$$H_0 : \theta^{(j)} = 0 \ , \qquad j = q+1,\dots,p$$

be the null hypothesis in the model (2). Denote by $\Lambda$ the likelihood ratio test statistic and define

$$\ell_{q,p} = 2(p-q)^{-1} \log \Lambda \ , \tag{9}$$

Then it is easy to see that

$$\ell_{q,p} = \alpha - (p-q)^{-1}(\mathrm{AIC}(p;\alpha) - \mathrm{AIC}(q;\alpha)) \ . \tag{10}$$

If we substitute the likelihood ratio test statistic $\ell_{q,p}$ by a robust version, namely

$$\ell_{q,p}^{rob} = 2(p-q)^{-1}(D(R)-D(F)) \ , \tag{11}$$

where $D(F)$ is the minimum value of $\sum_{i=1}^{n} \rho(r_{i;p})$ and $D(R)$ is the minimum value of $\sum_{i=1}^{n} \rho(r_{i;p})$ subject to $H_0$, the dispersion of the residuals under the full and reduced models respectively (see Schrader and Hettmansperger, 1980; Ronchetti, 1982), we obtain

$$\ell_{q,p}^{rob} = \alpha - (p-q)^{-1}(\mathrm{AICR}(p;\alpha,\rho) - \mathrm{AICR}(q;\alpha,\rho)) \ . \tag{12}$$

(12) is the natural counterpart of (10) when using robust estimators and test.

## 3. CHOICE OF THE PARAMETER $\alpha$

In this section we propose a choice for the parameter $\alpha$ in $AICR(p;\alpha,\rho_c)$. It is based on the following result due to Stone (1977).

The Akaike statistic $AIC(p;2)$ is asymptotically equivalent to

$$-2L_p + \mathrm{trace}(M_2^{-1}M_1) \; , \tag{13}$$

where $-M_2$ is the $(p \times p)$ matrix of the second derivatives (with respect to $\theta$) of the log-likelihood function and $M_1$ is the $(p \times p)$ matrix of the products of the first derivatives. Since $AICR(p;\alpha,\rho_c)$ can be viewed as the Akaike statistic computed under the least favorable errors' distribution $g_0$ (see (8)), we obtain

$$M_1 = E\psi_c^2 \cdot Exx^T$$

$$M_2 = E\psi_c' \cdot Exx^T$$

where $\psi_c(r) = d\rho/dr = r$ if $|r| < c$

$= c \cdot \mathrm{sign}(r)$ otherwise .

Thus, $2\ \mathrm{trace}(M_2^{-1}M_1) = 2(E\psi_c^2/E\psi_c')p$ and we propose to choose $\alpha=\alpha_c = 2E\psi_c^2/E\psi_c' < 2$. Note that $\alpha_\infty = 2$ and $AICR(p;\alpha_\infty,\rho_\infty) = AIC(p;2)$ which is the classical Akaike statistic under normality.

## Remark

Hampel obtains another choice for $\alpha$ "by adding the average decrease of $\sum_{i=1}^{n} \rho(r_i)$ and the average increase of the total mean square error of fit due to a superfluous parameter under normality" (Hampel, 1983). His choice for $\alpha$ is

$$\alpha = E\psi_c^2/E\psi_c' + E\psi_c^2/(E\psi_c')^2$$

that differs little from 2 for the usual values of $c$ (e.g. $c$ between 1.3 and 1.6).

## ACKNOWLEDGEMENTS

The author is grateful to Prof. F.R. Hampel for stimulating discussions. Partial support of ARO (Durham) contract #DAAG29-82-K-0178 is also gratefully acknowledged.

# REFERENCES

A' ike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory.* Academiai Kiado, Budapest, 267-81.

Atkinson, A.C. (1980). A note on the generalized information criterion for a choice of a model. *Biometrika* 67, 413-8.

Bhansali, R.J. and Downham, D.Y. (1977). Some properties of the order c' 'n autoregressive model selected by a generalization of Akaike's FPE cr erion. *Biometrika* 67, 547-51.

Hampel, F.R. (1974). The influence curve and its role in robust estim on. *J. Am. Statist. Assoc.* 69, 383-93.

Hampel, F.R. (1983). Some aspects of model choice in robust statistics. Proceedings of the 44th Session of ISI. To appear.

Huber, P.J. (1981). *Robust Statistics.* Wiley. New York.

Mallows, C.L. (1973). Some Comments on $C_p$ . *Technometrics* 15, 661-75.

Martin, R.D. (1980). Robust estimation of autoregressive models. *Directions in Time Series.* Inst. of Math. Statist., 228-62.

Ronchetti, E. (1982). Robust alternatives to the F-test for the linear model. *Probability and Statistical Inference.* Reidel, Dortrecht, 329-42.

Schrader, R.M. and Hettmansperger, T.P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika.* 67, 93-101.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J.R. Statist. Soc.* B 39, 44-7.